DOCUMENT RESUME

ED 222 556 TM 820 710

AUTHOR Haertel, Edward
TITLE Developing a Discrete Ability Profile Model for Mathematics Attainment. Final Report.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.; Stanford Univ., Calif.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE [81]
GRANT NIE-C-80-0003
NOTE 57p.; Tables are marginally legible due to small print. For related documents, see TM 820 707-712 and TM 820 716.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Cluster Analysis; Educational Assessment; Elementary Secondary Education; *Item Analysis; Mathematical Models; *Mathematics Achievement; National Surveys; *Skill Analysis; Test Construction
IDENTIFIERS *National Assessment of Educational Progress; *NIE ECS NAEP Item Development Project; Second Mathematics Assessment (1978)

ABSTRACT

National Assessment of Educational Progress (NAEP) Mathematics Assessment (1982) data were analyzed using latent class models to determine patterns of distinct skills required by different exercises and to estimate the pattern distributions. The populations were 9-, 13-, and 17-year-old examinees. Skills were assumed to be intermediate between objectives and report topics, such as "solving quadratic equations," and were treated as dichotomous - an examinee either did or did not possess the skill. At age 9 and 13, one assessment booklet was selected and nine clusters of three exercise parts were chosen. Twenty pairings of clusters yielded a 6-item set for analysis. At age 17, six exercise parts of apparently common skills were drawn from each of six booklets. Item clusters which could be collapsed and organized hierarchically were indicated by the latent classes. For each cluster, all analyses including it were examined together, yielding separate estimates of the proportion of examinees able to solve items in that cluster. The distributions of these estimates were an indication of the cluster's conformity to the assumption of skill dichotomies. Student mathematics skills at each age level are reported. Results of the use of NAEP data tapes are reported and improvements in the methodology are suggested. Primary type of information provided by the report: Results (Secondary Analysis). (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED222556

Developing A Discrete Ability Profile Model for Mathematics Attainment

FINAL REPORT

Edward Haertel

Stanford University

TM 820 710

The work upon which this publication is based is performed pursuant to Grant NIE-G-80-0003 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency.

Abstract

Data from the year 9 mathematics assessment were analyzed using latent class models to determine patterns of distinct skills required by different exercises, and to estimate the distributions of these skill patterns in age 9, 13, and 17 populations. At each of ages 9 and 13, one booklet was selected, and nine clusters of three exercise parts each were chosen from that booklet. Over twenty pairings of these three-item clusters yielded the six-item sets analyzed. At age 17, six exercise parts that appeared to require the same common skills were drawn from each of six different booklets for analysis.

At age 9, roughly 89 percent of the children could do simple addition problems correctly. Subtraction, counting, working with place values, common measuring units, and simple geometric concepts were available to roughly 77 percent of nine-year-olds. Only 41 percent could do more difficult computations, and only 53 percent some problems requiring use of a ruler. Skills at age 13 were less clearly distinguished. Number line problems and simple computations could be solved by 74 percent of thirteen-year-olds. The radical sign was interpretable to 66 percent, but only 30 percent could give decimal equivalents to common fractions. Skills for various algebra topics were available to 40 to 60 percent of children at this age. From 85 to 92 percent could do unit conversions and knew some geometry facts and concepts. The age 17 analyses showed 52 percent unable to solve any but the most elementary exercises requiring an understanding that letters may represent variable numerical quantities. Another eight percent could solve some such exercises by reasoning, but could not solve those requiring formal algebra training.

No problems were encountered with the NAEP tapes. Improvements in the methodology, as well as areas where further work is required, are reported.

TABLE OF CONTENTS

Abstract. i
Table of Contents ii
List of Tables and Figures. iii
Introduction. 1
Conceptualizing Mathematics Attainment. 3
Items as Skill Indicators 6
Carrying Out the Analyses 9
Strategy for Analysis 16
Using the NAEP Data Tapes 18
Results and Interpretation: Age 9. 20
Results and Interpretation: Age 13 28
Results and Interpretation: Age 17 37
Conclusions 45

LIST OF TABLES AND FIGURES

Table 1. Age 9 Estimates of Misclassification Probabilities by Item. . 26

Table 2. Age 9: Summary Across Runs by Cluster. 27

Table 3. Age 13 Estimates of Misclassification Probabilities by Item . 33

Table 4. Age 13: Summary Across Runs by Cluster 35

Table 5. Exercises Examined in Age 17 Analyses, p-values, and Latent Classes Required. 39

Figure 1. Age 9 Booklet 3 Stem-and-Leaf of Crude Item Difficulties . 21

Figure 2. Age 9 Analyses and Proportions in Each Skill Pattern 23

Figure 3. Age 13 Booklet 6 Stem-and-Leaf of Crude Item Difficulties . 28

Figure 4. Age 13 Analyses and Proportions in each Skill Pattern. . . . 31

DRAFT FINAL REPORT: Developing a Discrete Ability Profile Model for Mathematics Attainment

The National Assessment of Educational Progress affords a rich resource for the description of American young people's academic skills. Careful deliberation as to appropriate objectives, matching of item formats to objectives, including the use of unconventional formats as necessary, and use of all (technically sound) exercises regardless of their "item statistics" yield item pools that tap an unparalleled array of specific competencies. In contrast, traditional tests sample a narrower range of objectives, often constrain all items to a common format, and typically include only items that are highly correlated with total score or some other, single criterion. Matrix sampling permits the administration of literally hundreds of exercises, all to nationally representative respondent samples, without unduly burdening any individual pupils or schools.

Unfortunately, researchers seeking to exploit these data to find out what children have learned or can do may be overwhelmed by the embarrassment of riches. Accustomed to conceiving items as multiple indicators of a single ability, they are ill-equipped to draw useful conclusions from hundred of diverse items, each of interest in its own right. The most typical response has been to attempt to build scales, using items that

appear to be related, and restricting attention to one exercise booklet (package) at a time.

Those turning to NAEP publications of findings will find little additional guidance. Here, results are aggregated over packages, but exercises are typically pooled into broad "report topics" categories spanning many objectives, simply because objectives are so numerous. The only direct longitudinal comparisons made are of performance on identical sets of items at different points in time, and for descriptions of absolute level of performance at a single time point, the reader is told, "30% of nine-year-olds know this...", "45% can solve this problem," etc. It is then left to the reader to imagine what generalizations from these statements are or are not appropriate. Only 7% of seventeen-year-olds could correctly solve the equation "$(x-2)^2 = 9$" for x, but in another sample, 18% could "Find the solution set of $x^2 - 5x + 6 = 0$."* Is the difference due to the wording of the problems? The particular numbers? The format of the equations? How are we to generalize about the proportion of seventeen-year-olds who can solve quadratic equations? How would the p-values change if these items were multiple choice, say, rather than free response? There is no way to tell.

Here is a dilemma. The very breadth and richness of the data base make it resistant to all but the most superficial summaries! The conventional model of items as multiple indicators of a single ability continuum is not appropriate, and there appears to be no alternative but to compose average p-values for masses of exercises. There is clearly a need for better methods of reporting NAEP findings. The purpose of this study was to investigate one possible method.

*Exercises S0429A and S0905A from year 09 mathematics assessment.

Conceptualizing Mathematics Attainment

An empirically grounded summary of NAEP findings must be based on some model for exercise response data. Traditional models, both classical and latent trait, are most often appropriate when (1) there are large numbers of items that may be assumed unidimensional, and (2) the intent is to make inferences about individual examinees' levels of the ability underlying responses to all the items. In contrast, the NAEP data offer (1) only a few exercises tapping each of many abilities and (2) a problem of describing performance aggregated across individuals. Clearly, then, a different kind of model is needed.

For this study, the kind of model chosen was a restricted latent class model similar to those first investigated by Lazarsfeld and Henry (1968), and applied to item response data by Proctor (1970), Dayton and Macready (1976), Goodman (1975), Haertel (1980), and others. The basic "unit" of ability assumed in these models may be termed a "skill." For purposes of this study, skills were assumed to be intermediate in scope between objectives and report topics. For example, "solving quadratic equations," or "making conversions among common units of measurement" might be skills. A distinctive feature of the models is that skills are treated as dichotomous: A given examinee either does or does not possess the skill. If an item requires only skills an examinee possesses, then s/he can solve that item. If it requires even one additional skill, s/he cannot. When applied to individual-level data where the intent is to distinguish a continuous range of ability levels, this might not be a useful assumption. For describing populations, however, it works well. With respect to any given skill dichotomy, examinees can be partitioned (conceptually) into

just two groups: those that possess the skill and those that do not. The proportion in the first category is sufficient to summarize completely the distribution of that skill in the population.

When multiple skills are considered, it may be that hierarchies will be found. For example, it is unlikely that any students possess the skill of solving quadratic equations but lack the skill of solving linear equations in one unknown. Thus, these two skills together define only three groups of examinees: those possessing neither, only the linear, or both. On the other hand, students who can do conversions among common units of measurement may or may not be able to solve linear equations, and conversely. These two skills define four groups of examinees: those possessing neither, only the first, only the second, or both. The set of all skills defines a very large number of possible skill profiles, and under these models, a list of the proportions of examinees in each possible profile would completely describe the distribution of mathematics abilities in the population. While such a complete cross-tabulation of skills possessed or not possessed is not obtainable in practice, various marginal tabulations involving subsets of the skills can be estimated, and provide useful generalizations across items.

Before explicating details of the models, estimation procedures, tests of fit, etc., it will be useful to consider more carefully the reasonableness of the dichotomy assumption, and sources of skill distinctions in the population. When a single examinee encounters a single exercise, it seems reasonable to suppose that s/he either can or cannot get it right. If the exercise could be given the same child repeatedly with perfect forgetting between administrations, we would expect either nearly all incorrect responses or nearly all correct responses, with the imperfections due only to guessing or to carelessness. Thus, it does seem reasonable that a single item defines a dichotomy, and not a continuum.

After all, responses are scored dichotomously, so an item in isolation can give us no more information than "probably present" or "probably absent" concerning whatever skills it requires. Next, if two items are considered, each will define such a binary partition. Examinees will be able or unable to solve the first, and able or unable to solve the second. If the items are quite similar in content, it may be a good approximation to suppose that they in fact define the same latent dichotomy, and that examinees who get one right and the other wrong have guessed a right answer or else have been careless in giving a wrong one. In fact, if we are free to estimate the probabilities of "false positives" and "false negatives" for the two items as well as the proportion who really can solve the two items, such a model can always reproduce perfectly the actual frequencies of each possible response pattern to two, or even three, items. With from four to six items that are homogeneous in content, this model, with its single latent dichotomy, is virtually indistinguishable from continuous unidimensional latent trait models in its ability to reproduce actual frequencies of different patterns of item responses.* Obviously, with a large pool of items testing the same complex ability, we can distinguish more than two levels of ability in examinees. The point is that with fewer and/or more homogeneous items the alternative assumption, that examinees either have or have not acquired the skill, may be as good or better for some purposes.

*Findings are from unpublished investigations by the author. Comparisons were made between the two-state latent class model and the two-parameter normal ogive model (Bock & Lieberman, 1970), which has an almost identical number of parameters.

The skills assumed to be measured by NAEP exercises are not basic psychological processes, but arise largely from the organization of the mathematics curriculum in American schools. Educators have chosen to parcel out the content of mathematics into courses such as Algebra I, Geometry, etc. Topics usually taught together are likely to cohere in single skills, but if the organization of the curriculum were changed, the skills would change, too. Of course, if some topics require that students attain a given developmental level before they can be acquired, then skill patterns involving those topics will reflect cognitive processing capabilities as well as curricular conventions. No attempt was made in this study to disentangle sources of skill distinctions. The objectives were to determine what skills need be assumed, and how prevalent they are in the population.

Items as Skill Indicators

In the latent class models used in this study, each exercise is assumed to require one or more distinct skills, and examinees are assumed to either possess or not possess each skill. The skills an exercise requires and an examinee possesses are the sole determinants of the probability that the examinee will answer correctly. There will be one (low) probability of a correct response by an examinee lacking one or more requisite skills, and a second (high) probability of a correct response by an examinee possessing all the skills the exercise requires. Responses to different exercise parts are assumed conditionally independent given skill possession. Thus, for any single examinee or for any group of examinees who possess the same

skills, the probability of a pattern of responses across exercises is just the product of the separate probabilities of each component response. It is this assumption of conditional independence that makes it possible to estimate parameters of the latent class model. An exactly analogous assumption is required in all commonly used latent trait models for item response data, as well.

It follows from these assumptions that a given exercise can be completely characterized by telling (1) what skills it requires, (2) its false positive rate, i.e., the (hopefully low) probability that any examinee lacking one or more requisite skills will answer it correctly, and (3) its true positive rate, i.e., the (hopefully high) probability of a correct response by any examinee possessng all the requisite skills. The hypothesized skill requirements of each of a set of exercise parts determine what latent classes are assumed when that set is analyzed. Following the analysis, a fit statistic and residuals are examined, and the hypothesized skill requirements are revised as necessary to obtain a satisfactory fit to observed response pattern proportions.* The analysis consists of estimating simultaneously the proportions of examinees in each assumed latent class, and the false positive and true positive response probabilities (FP and TP rates) for each exercise part. The p-value for each exercise part is a weighted average of its FP and TP rates, the weights being the total proportions of examinees who should be unable to solve and

*Observed response pattern proportions are the fractions of examinees who each give possible pattern of correct and incorrect responses to the set of exercise parts being analyzed.

able to solve that part, respectively.

The estimated FP and TP rates are assumed to be parameters of the item itself, and in fact are found to be quite stable across estimations where the item is included in different sets. FP rates indicate susceptability to guessing. They are typically higher for multiple choice than for free response items, but may also be high for items that invite some alternative solution that circumvents the intended skill requirement. For example, one free-response item presents the stimulus

$$\square - 19 = 32$$

and asks, "what number should go in the $\square$ to make this number sentence TRUE?"* When this item is coded as requiring the same skill as two other items that require a basic understanding of equations, a satisfactory fit is obtained (about 63% of thirteen-year-olds possess the skill), but the false positive rate is estimated at .22. This reflects not guessing in the usual sense, but the possibility of obtaining the correct answer without understanding, by one of the few possible elementary operations that can be performed using the numbers shown -- the answer is 19 + 32.

The TP rate should reflect primarily carelessness in responding, but also can indicate idiosyncratic informational requirements or unique difficulties of single exercise parts. For example, a skill of converting units was assumed to underlie three (unreleased) multiple choice items requiring conversions between quarts and gallons, ounces and pounds or feet and yards, respectively.** While models assuming a single skill did fit

*Exercise T0608A from the year 09 mathematics assessment.

**T0616B, T0616C, and T0616D from the year 09 mathematics assessment.

these items (roughly 53% of thirteen-year-olds possess the skill), their estimated true positive rates were on the order of .84, .90, and .87, respectively. Thus, 16% (1 - .84) of those who possessed the skill lacked specific information on quarts and gallons, 10% lacked specific information on ounces and pounds, etc.

The difference between an item's TP and FP rates may be interpreted as an index of item discrimination.* A large difference indicates that the item reliably distinguishes between those who do and do not possess the skills required for its solution, while a small difference indicates that correct responses by nonmasters and masters are more nearly equally likely.

Carrying Out the Analyses

This section describes the computational steps required in applying latent class models with the NAEP data. The overall approach taken in studying mathematics skills, of course, includes more than these technical procedures. In the next section, "Strategy for Analysis", substantive decisions are described, including questions as to which booklets to examine, which items to group together, and how to integrate findings from many separate analyses to reach meaningful conclusions. Before turning to these broader issues, details of the methodology are described below.

*An optimal discrimination index is the log odds ratio, $\ln \frac{TP(1-FP)}{FP(1-TP)}$, which can be used in exactly the same way as the discrimination parameter of the two-parameter logistic model in constructing an optimum scoring rule.

The latent class analysis begins with a summary of the data that tells what proportion of examinees gave each possible pattern of correct and incorrect responses to a set of items. There are both advantages and disadvantages to this approach. On the one hand, the costs of computation are little affected by sample size, since response pattern proportions may be obtained in a single pass over the data file, and need not be re-calculated each time a new model is fit to the same items. The number of possible patterns, hence the number of proportions, is independent of the N. In addition, issues of weighting and other adjustments for the NAEP sampling design may be handled prior to the actual analysis. On the other hand, summarizing the data according to response pattern proportions severely limits the number of items or exercises that can be examined at one time. The number of possible response patterns is four for two items (00, 01, 10, 11), eight for three items (000, 001, 010, 011, 100, 101, 110, 111), sixteen for four items, and in general, 2^k for k items. In examining the NAEP data, sets of six exercise parts were employed. Thus, data for all respondents could be summarized in just 64 numbers, corresponding to all possible patterns of correct and incorrect responses, from 000000 to 111111. Since the intent of this study was to distinguish the specific skills required by individual exercise parts, sets of six at a time were quite sufficient. However, in order to attain a clear conception of the skills assessed by the entire exercise pool, it was necessary to integrate findings across many separate sets of items.

After obtaining response pattern proportions, the next step was to fit a latent class model to these proportions, in order to determine which of the exercise parts could be assumed to require the same skills and which must be assumed to require different skills. In each single computer run,

just one model is tested. Based on the fit of this model, it is either accepted as accounting adequately for the data obtained, or rejected. In the latter case, information on just where and how the model failed to fit (analysis of residuals) is used to guide the selection of a new model to try, and the cycle is repeated until an acceptable solution is found. It is for this final model that item misclassification parameter estimates and latent class proportions (estimated proportions of respondents possessing each combination of skills) are most carefully examined.

Specific steps of the procedure are as follows. After choosing a set of six exercise parts to be analyzed, a table is prepared showing the location of each part in the file and its correct response value. The original Fortran program used to access the tape requires a one-column numeric field for each item, so for free-response items the first column of the (typical) two-column field was examined for a "1". Control cards for the program, called PULL, specify logical record length, column numbers, and keys for each item, and the first and last column of the weight variable, if used. Additional information as to which tape file to access and the name to be given to the output data set is coded in the JCL for the PULL run. For all analyses, responses were weighted using the variable WEIGHTS, "student weight." The PULL program read each record, scored the items to determine the response pattern, and incremented a tally for that pattern by the weight for that record. Thus, the total weight for each pattern was obtained. These were then divided by the grand total weight to obtain weighted response pattern proportions, which were written to a small data file along with the original grand total weight. The PULL program could actually process up to six item sets at a time, producing six of these small data files from a single pass over the data. The cost of a typical run to produce six data sets for a single exercise package at normal daytime rates was $3.32, including a $2.00 charge for mounting the

tapes.*

The small files were next edited and copied into a "library" file for storage. Since each could be contained on only 12 80-column cards, it was most efficient to keep them online. The editing required was to replace the total weight with the actual N, adjusted according to design effect for the sample. What was done was to set the N equal to one half the number of records in the file, i.e., to assume a design effect of 2.00. While this adjustment cannot be justified rigorously, various indirect tests have indicated that it is satisfactory in practice (Haertel, 1980).

The next logical step in the analysis was to input the response pattern proportions to a second Fortran program**, MLMN, which produces maximum likelihood estimates of latent class proportions and misclassification

*In one case, a slightly more elaborate procedure was used, in order to combine four exercise parts into a single overall exercise score. An additional Fortran program was used to read each record from the tape, score that one exercise, and write the record to a temporary disk file. The disk file was then input to the PULL program. This was done for exercise parts T0632A, T0632B, T0632C, and T0632D from the year 09 mathematics assessment.

**This program was written by Richard Wolfe, currently at the Ontario Institute for Studies in Education. It has been modified by the Principal Investigator for use in this study, but is not generally available. However, newer programs incorporating more recently developed estimation algorithms are generally available and can be used to carry out analyses like those reported. The principal advantages of the MLMN program are superior numerical accuracy, calculation of asymptotic standard errors and covariances of the estimates, and detailed analyses of residuals useful in deciding how to revise provisional models to improve the fit.

probabilities, together with associated statistics. However, because the control cards required for MLMN are quite complex, another Fortran program, PREP, is executed first, to prepare the stream of control cards for MLMN. The PREP program needs very brief control cards telling which latent classes are to be included and giving labels for the model, the six-item set, and each separate exercise part. Output from PREP is a jobstream which can be input to MLMN.

The first run for each set was to fit the simplest possible latent class model, with only two latent states. These are labeled the "null" state and the "all" state, and are included in any subsequent models, as well. It is assumed that examinees conforming to the "null" state cannot solve any of the six exercise parts, so any correct response by these examinees must be a false positive. Examinees conforming to the "all" state are assumed capable of solving all items correctly, so that any correct responses they provide are true positives, and any incorrect responses are false negatives.

MLMN estimates the proportions of examinees in each latent class, and probabilities of correct responses to each item by members of each class. Fit is assessed by both likelihood ratio and Pearson chi square statistics. If the fit of this initial model is satisfactory (as indicated by a non-significant chi square), it indicates that the assumption of a single underlying skill dichotomy common to all six items cannot be disconfirmed. In this case, analysis of this exercise set is completed. Because sets of exercises are chosen to detect distinct skills, however, it more often happens that the chi squares for this initial model are unacceptably large.

In this case the analysis of residuals is consulted to determine what additional latent classes may be required to improve the fit. It should be explained that the parameter estimates are used to predict the proportions of examinees expected for each response pattern. These fitted

(predicted) values are subtracted from the observed values to obtain residuals. If the model fits the data, residuals are expected to be small and non-systematic. If the model fitted does not adequately account for the observed response pattern proportions, systematic patterns in the residuals usually offer clues to more complex models that would do a better job of accounting for the data. The problem in analyzing residuals is to detect these patterns.

Several methods for analysis of residuals have been tried by the author, and the problem is by no means solved. The best method thus far is as follows. Consider the residuals for the 64 response patterns to be entries in a 2 x 2 x 2 x 2 x 2 x 2 table, with each dimension corresponding to an item, and the two levels of the dimension corresponding to incorrect ("0") or correct ("1") responses to that item. Then, obtain all possible marginals of the table by summing across every possible combination of one or more dimensions. For each table formed by collapsing across just one dimension, there will be 32 marginals. When two dimensions are collapsed there will be 16 marginals, etc. Of these, list, for each possible collapsing of the original table, the one marginal for "correct" responses to all remaining items. When one of these values is large, it suggests that the items not collapsed in obtaining that value covary to a greater extent than the current model permits. Thus, a new latent class corresponding to examinees who can solve those items but not the remaining items is called for. In other words, the analysis of residuals permits identification of subsets of exercise parts that require a common skill, not required by the remaining exercise parts. When the corresponding latent class is introduced, the fit is generally improved.

After the model is revised and a new run is made, a <u>difference chi square</u> is calculated. This is simply the difference between the original

and new likelihood ratio chi squares, and is asymptotically distributed as chi square on one degree of freedom, if only one new parameter was introduced and if the less inclusive model fits the data. A significant difference chi square indicates that the additional latent class should be retained; if the overall fit is still not satisfactory, additional latent classes must be introduced, based on analysis of residuals from the new model. If the difference chi square is not significant, the new latent class is not retained in the model, and a different latent class is introduced instead. This procedure is repeated until a satisfactory fit is obtained, no additional latent classes can be found that yield further improvements in fit, or the estimated proportions of examinees in successive states fall below about 2%. These latter situations are rare; in almost all cases, the chi square becomes acceptably small with no more than four latent classes, including the "null" and "all" classes. Acceptable fit is defined by a chi square below 70 (usually below 65). Because the design effect adjustment directly affects the value of the chi square*, rigid adherence to a precise significance level in interpreting chi squares is not appropriate. Chi squares of 65 to 70 on 48 to 50 degrees of freedom are significant at roughly the .05 level.

*Adjusting for a design effect of 2 will halve the value of both the likelihood ratio and the Pearson chi squares, will increase all standard errors by a factor of $\sqrt{2}$, but will not affect the magnitude of any parameter estimates, fitted response pattern proportions, or residuals.

Strategy for Analysis

This section describes the use of latent class models in examining skill distinctions in mathematics attainment. The first step for each age level was to choose a representative exercise package (booklet). The appendices were dumped from tape, and Appendix 4 was used to find the booklet with the best representation across the areas of algebra, geometry, measurement, and computation.* Microfiche of the actual booklets was also consulted in making this determination. The codebook file for the selected booklet was then dumped, and hard copy was prepared from the microfiche for the booklet selected.

On the basis of Appendix 4 classifications as well as direct inspections of the exercises, all cognitive exercise parts were classified as Computation, Algebra, Geometry, Measurement, or Other. ("Numbers and Numeration" exercises were classified under computation.) Within each category, a stem-and-leaf of "crude item difficulty" was prepared. These "crude item difficulties" were the (unweighted) proportions of correct responses reported in the codebook.

Under the assumptions of the latent class models, there is, of course, no necessity that items requiring the same skills be of similar difficulty. The item difficulty reflects not only the proportion of examinees possessing requisite skills, but also each item's unique misclassification

*If in future assessments exercises are packaged to provide topical coverage (i.e., if different booklets focus on different themes), it would be helpful for this kind of analysis if at least one or two booklets were organized according to the present practice, with a few exercises of each of many types.

probabilities, reflected in its item parameters (true positive rate and false positive rate). Nonetheless, it is more likely that items similar in difficulty will require the same skills than test items of widely differing difficulties. Thus, the stem-and-leafs were useful in identifying initial item clusters. Exercise parts within a category and with similar crude difficulties were examined, and subsets of three were chosen that appeared relatively homogeneous in content. In some cases, these were three parts of a single exercise. In other cases, two or three distinct exercises were represented. These triples, similar in content and difficulty, were the units used to form the sets of six exercise parts analyzed at ages 9 and 13. (For age cycle 17, a different procedure was followed.) Thus, each analysis at ages 9 and 13 involved exactly two three-item clusters. A two-class model was fit to the six exercise parts, and if satisfactory fit was not obtained, additional states were introduced based on analysis of the residuals. For all but 3 of the 47 sets analyzed at ages 9 and 13, the additional skill distinctions required distinguished between the two clusters but not within clusters. In other words, the three exercise parts within each cluster required identical skills.

Nine 3-item clusters were identified at each of ages 9 and 13. Thus, a maximum of 36 runs at each age were possible, taking all cluster pairs. Resource limitations prevented carrying out all these analyses. (Recall that one "analysis" could require as many as six or more runs.) Twenty-six analyses at age 9 and twenty-one at age 13 were selected. In order to locate all six-item sets for which a single dichotomous skill could be assumed (sets for which the two-class model fit), or for which skills would be hierarchical (one triple requiring a subset of those skills the other required, as indicated by an acceptable fit for a three-class model), adjacent clusters within a content strand were paired first. Additional analyses were then included which involved clusters of similar difficulty

in different content areas.

After all analyses were performed, results were summarized in figures and tables, and examined to determine which, if any, item clusters could be collapsed and which could be organized hierarchically. This was indicated by the latent classes required in each run. Next, for each cluster, all analyses including that cluster were examined together. Each of these analyses yielded a separate estimate of the proportion of examinees able to solve items in that cluster. The distribution of these estimates is one indicator of the cluster's conformity to the assumption of skill dichotomies. The separate estimates of each item's true positive and false positive rates were also tabled.

Many exercises in each chosen booklet were excluded from the analyses. A basic framework was developed to encompass most exercises at each age level, but available resources did not permit the additional analyses necessary to include all the exercise parts in the booklet. Extension of the basic structure is straightforward logically, but costly in human and computer time, as additional content clusters must be paired with increasing numbers of clusters already identified.

Using the NAEP Data Tapes

The tapes and documentation provided by the National Assessment were outstanding. Compared to other large data sets I have worked with, there were virtually no problems with either tapes or documentation. All but one or two of the runs directly accessing the tapes ran correctly the first time they were submitted.

IBM standard-label tapes were provided by ECS. Machine-readable documentation files (fixed-length records with ASCII control characters) were dumped using IOPROGM, a locally-written utility program. Data were

read using formatted READ statements in an original Fortran program. Record layouts were easy to follow and appeared absolutely accurate. Weight variables were well-documented as well. I found it especially useful to have the unweighted frequency distributions of each variable in the Codebook files. My use of these distributions was described earlier.

The few improvements I suggest below are minor; I was satisfied with the tapes exactly as provided.

- It would be helpful to have appendices and codebook files on microfiche as well as on tape. As it is, it may be necessary to dump thousands of lines to answer a single, simple question.
- There is a good deal of redundancy in appendices for different age levels. If appendices were in separate files and if redundant files were identified in the printed documentation sent by ECS, some needless printing could be avoided.
- Some structure and consistency in the data is not documented in such a way that the user can take advantage of it. For example, the WEIGHTS variable is in the same columns in all booklets I examined at all age levels, but I did not find any summary indicating what portion of the data records is fixed across booklets and ages.
- For many purposes, it would be useful to be able to construct interpenetrating subsamples within each booklet, e.g., jackknifing, crossvalidation, or replication of analyses to estimate standard errors. Construction of interpenetrating subsamples would be facilitated if a one-column "subsample ID" were included on

each record, to be used in allocating records to separate subfiles. Within strata, PSUs would be randomly allocated to subsamples. Records within PSUs would not be divided, to avoid any possible error covariation between samples

Results and Interpretation: Age 9

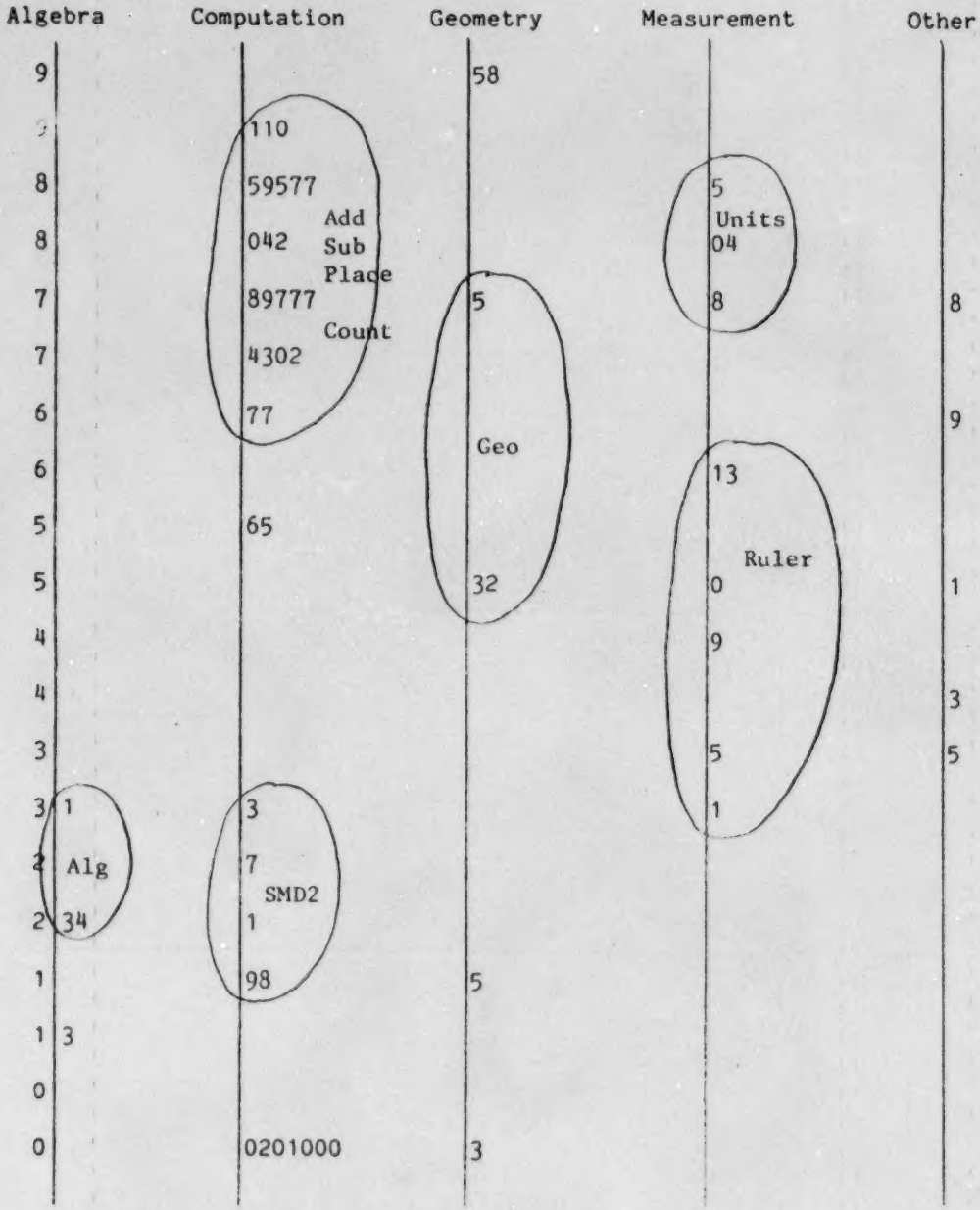
Figure 1 displays stem-and-leafs of crude item difficulties for all exercise parts in Booklet 3, Age 9. Homogeneous sets of exercise parts have been circled, and the clusters (triples) drawn from each set are indicated. The three "Alg," or Algebra items include two using a number line and one asking which number sentence would be used to solve an equation. All three were multiple-choice exercises. The "Add" and "Sub" clusters include simple addition and subtraction problems presented orally. These were free-response. "Count" included counting sets of small illustrations in the exercise booklet, by ones, by twos, and by tens. The first of these, counting squares by ones, was free response. The latter two, counting shoes in pairs and marbles in bags of ten, were multiple-choice. "Place" items included giving the place values of specific digits in numbers, and an exercise like writing "ten sevens".* All were multiple-choice. The more difficult computation clusters, "SMD2," included two- or three-digit subtraction, multiplication, and division problems. These were similar to* 47 x 6, 605 - 328, and 56 ÷ 4, all free-response. "Geo" (geometry) included multiple-choice items asking in which figure the halves would not match if it were folded, which figure illustrated parallel lines, and which illustrated three line segments that could not make a triangle. The two-choice "units" items asked which is more: a yard or two feet, two quarts or three pints, two dimes or three nickels. Finally, the free-response "Ruler" items involve using a ruler to measure a line,

*Actual content of secure exercises has been modified to avoid disclosure.

measure the distance around a triangle, and draw a line of a specified length. (Exercise numbers assigned to each of these items in Booklet 3 are shown in Table 1, discussed below.)

Figure 1

Age 9 Booklet 3 Stem-and-leaf of Crude Item Difficulties



The 26 analyses run on these clusters are diagrammed in Figure 2. Each line connecting two clusters represents one analysis. The analysis number (1-26) is written along the line, along with three proportions separated by tick marks, which summarize the skill distinctions found in that analysis. The center proportion indicates what fraction of examinees were able to solve all exercises in both clusters. The proportions on the ends closest to each cluster give the proportion able to solve items in that cluster but not the other. If one of these is 0, a hierarchical relationship is indicated. No examinees are able to solve those items who cannot also solve the items in the other cluster. For example, Analysis 7, involving the Geo and SMD2 clusters, yielded the proportions .158/.444/0. This indicates that 15.8% of the examinees could solve Geo but not SMD2 items, 44.4% could solve either type, but none were capable of solving the SMD2 items and not the Geo items. Thus, Geo items fall below SMD2 items in a hierarchy. Note that the remaining 39.8% of the examinees (100% - 15.8% - 44.4%) could not solve any of the six exercise parts in this analysis.

In some cases, two of the three estimated values were zero, as in Analysis 11 involving the SMD2 and Alg clusters. In this case a single, common skill distinction underlying exercises in both clusters proved sufficient to explain the data. The two clusters collapsed into a single set requiring the same, common skill.

For only two analyses at age 9, analyses 6 and 17, were items within a cluster found to require distinct skills. In both cases, the Place exercise requiring the examinee to select something like "ten sevens" migrated to the other cluster. In analysis 6, this item proved to have more in common with counting by ones, twos, and tens than with the other Place items. In analysis 17, the same item migrated to the Geo cluster.

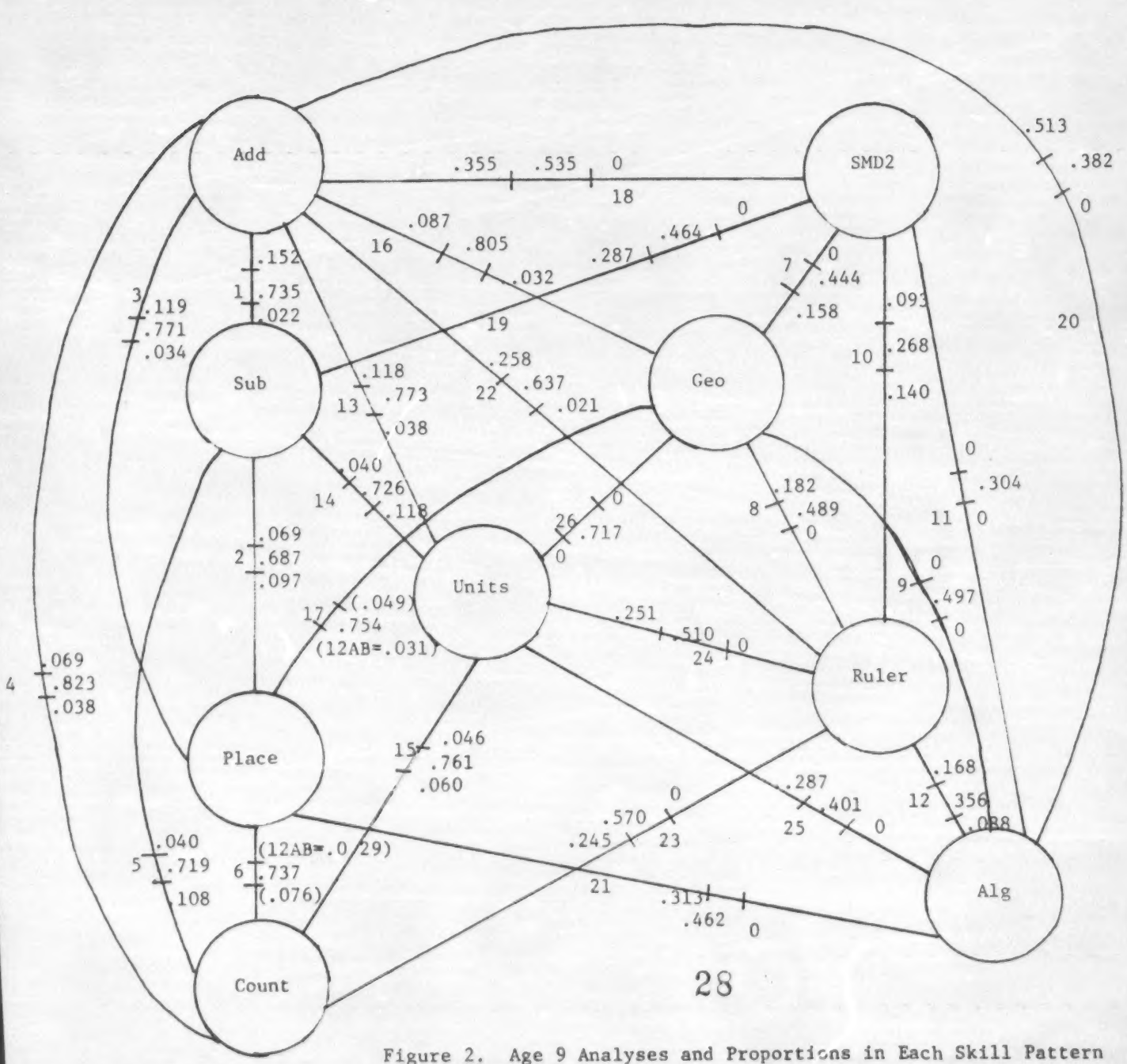


Figure 2. Age 9 Analyses and Proportions in Each Skill Pattern

Figure 2 reveals a larger number of distinct skills than expected. Only two pairs of clusters collapsed, leaving seven distinct skills. Clusters collapsing were Units with Geo, and SMD2 with Alg. With only one minor exception, the 26 runs indicated the same hierarchical pattern among these seven skills: At the first level are the skills Add, Sub, Count, Place, and Units-Geo. These may all be acquired independently, although possession of Sub and not Add is quite rare (2.2% of the population), as are possession of Count and not Add (3.8%). Place and not Add (3.4%) or Units-Geo and not Add (3.5%). At the second level of the hierarchy are two skills that may be acquired independent of one another, but are never found in the absence of any of the first five skills. These are SMD2-Alg and Ruler. The only minor exception is in analysis 22, which indicates 2.1% able to solve Ruler items but not Add items. These hierarchical dependencies may stem from two sources. First, SMD2 and Alg items may require as component operations the skills entailed in Add, Sub, Place, etc. Thus, the structure of the items gives rise to a logical dependency. Second, elementary mathematics curricula may be so organized that the Alg-SMD2 and Ruler skills are almost never introduced until after the skills at the first level of the hierarchy. This would result in very few students being able to solve the former and not the latter.

The next step in interpreting and integrating results of the analyses was to table the item parameter estimates, and estimates of proportions able to solve each item type. This tabulation is displayed in Table 1. The leftmost columns of Table 1 give cluster and item identifications, followed by item difficulties (p-values) calculated weighting each response according

to the WEIGHTS variable. The remaining columns give identifying information and results for each analysis in which the cluster was included. The TP and FP columns for each analysis present true positive and false positive rate estimates for each item. To the left of these estimates is a column of identifying information and results giving the analysis number and the other set of items included in that analysis (line 1), the chi-square for the final model (line 2), and the total proportion able to solve the cluster, as estimated in that analysis (line 3).

As can be seen in Table 1, estimates of the proportion able to solve the cluster were extremely stable for the clusters Add, Sub, Place, and Count. A wider range of estimates was obtained for Units, SMD2, Ruler, and Alg, and estimates for Geo were fairly unstable. Consistency of estimates appears to be related to the homogeneity of the exercises in the cluster. For the clusters that collapsed, Units - Geo and SMD2 - Alg, estimates of the proportion possessing the common skill were consistent across the two clusters. Averages were .42 and .41 for SMD2 and Alg, respectively, and .69 and .77 for Geo and Units. Note that the discrepancy of .08 between Geo and Units is not large given their standard errors (on the order of .025 for most estimates, yielding a t of 2.26) and is much smaller than the difference between mean p-values for items in the two clusters (.63 for Geo and .81 for Units).

Estimates of proportions able to solve each cluster are summarized in Table 2. Note that these estimates are not statistically independent. Table 2 is purely descriptive. The estimates for each cluster are not representative of any examinee or item population, and reported standard

Table 1. Age 9 Estimates of Misclassification Probabilities by Item

luster	Item	P	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP
Add	27B	.914	01:Sub	.98	.38	03:Place	.98	.37	04:Count	.98	.36	13:Units	.98	.37	16:Geo	.98	.37	18SMD2	.98	.37	20:Alg	.98	.35	22:Ruler	.98	.35
	27D	.863	97.39	.96	.11	60.46	.96	.09	65.49	.96	.10	38.97	.96	.08	49.54	.96	.08	69.21	.96	.09	36.78	.96	.08	52.81	.96	.07
	27F	.880	.89	.97	.17	.89	.97	.16	.89	.97	.14	.89	.97	.16	.89	.97	.16	.89	.97	.15	.89	.97	.13	.90	.97	.13
Sub	16B	.844	01:Add	.97	.41	02:Place	.97	.41	05:Count	.97	.40	14:Units	.97	.39	19:SMD2	.97	.43									
	16D	.685	97.39	.88	.09	93.08	.88	.07	59.03	.88	.07	40.70	.87	.07	79.76	.88	.09									
	16F	.721	.76	.92	.10	.76	.92	.12	.76	.92	.11	.77	.91	.09	.75	.92	.12									
Place	12A	.796	02:Sub	.95	.23	03:Add	.94	.20	06:Count	.97	.24	17:Geo	.96	.21	21:Alg	.96	.24									
	12B	.859	93.08	.98	.41	60.46	.98	.36	62.83	.99	.42	50.60	.99	.40	40.65	.93	.42									
	29A	.791	.78	.89	.42	.80	.88	.41	.78* (.77,.77,.81)	.90	.32	.79* (.79,.79,.80)	.90	.34	.78	.89	.45									
Count	13A	.925	04:Add	.97	.67	05:Sub	.97	.70	06:Place	.96	.76	15:Units	.97	.72	23:Ruler	.97	.74									
	13B	.802	65.49	.90	.22	59.03	.91	.30	62.83	.92	.38	42.98	.91	.32	67.89	.91	.31									
	13C	.920	.86	.90	.35	.83	.90	.42	.81	.92	.40	.82	.92	.38	.81	.91	.40									
Units	05A	.788	13:Add	.87	.44	14:Sub	.87	.35	15:Count	.88	.39	24:Ruler	.91	.44	25:Alg	.92	.49	26:Geo	.90	.51						
	05C	.808	38.97	.92	.35	40.70	.89	.35	42.98	.90	.43	49.49	.92	.48	38.79	.94	.52	36.88	.94	.49						
	05E	.846	.81	.89	.66	.84	.88	.64	.81	.89	.65	.76	.90	.70	.69	.90	.72	.72	.89	.73						
Geo	14A	.771	07:SMD2	.89	.58	08:Ruler	.88	.56	09:Alg	.90	.64	16:Add	.84	.40	17:Place	.85	.46	26:Units	.86	.54						
	28A	.568	52.01	.71	.35	70.67	.70	.30	49.97	.73	.40	49.54	.64	.22	50.60	.63	.30	36.88	.65	.35						
	30A	.536	.60	.63	.39	.67	.62	.37	.50	.65	.42	.84	.59	.25	.80	.59	.32	.72	.60	.37						
SMD2	06A	.340	07:Geo	.61	.12	10:Ruler	.68	.15	11:Alg	.69	.19	18:Add	.59	.05	19:Sub	.60	.10									
	08A	.278	52.01	.58	.03	69.22	.63	.08	47.30	.69	.10	69.21	.51	.00	79.76	.57	.02									
	21B	.187	.45	.36	.05	.36	.41	.06	.30	.45	.07	.54	.32	.04	.46	.34	.05									
Ruler	37A	.619	08:Geo	.86	.39	10:SMD2	.86	.45	12:Alg	.85	.36	22:Add	.78	.31	23:Count	.82	.35	24:Units	.85	.38						
	38A	.367	70.67	.65	.10	69.22	.73	.12	45.57	.62	.09	52.81	.56	.00	67.89	.60	.06	49.49	.64	.08						
	39A	.659	.49	.91	.42	.41	.94	.47	.52	.90	.39	.66	.84	.30	.57	.88	.36	.51	.90	.41						
Alg	18A	.324	09:Geo	.50	.15	11:SMD2	.48	.25	12:Ruler	.55	.14	20:Add	.50	.19	21:Place	.53	.15	25:Units	.57	.16						
	31A	.234	49.97	.33	.14	47.30	.31	.20	45.57	.32	.17	36.78	.35	.16	40.65	.31	.16	38.79	.33	.17						
	34A	.240	.50	.39	.09	.30	.51	.12	.44	.43	.09	.38	.47	.10	.46	.43	.07	.40	.46	.09						

*Average proportion for 3 items. In these runs, the three place items showed distinct skill profiles.

deviations should not be used to construct standard errors or confidence intervals. For convenience, observed item p-values are also reproduced in Table 2. It may be noted that for multiple-choice exercise clusters Place, Count, and Units, observed p-values were inflated due to guessing. For the other two multiple-choice clusters, Geo and Alg, lower true positive rates offset the higher guessing probabilities, and true proportions able to solve these items exceeded the average p-values for the clusters.

In summary, the Age 9 analyses revealed seven distinct skills underlying the 27 exercise parts examined, two of which were acquired only after all of the remaining five. While the latter five skills could be acquired in any order, only a few examinees lacking the Add skill possessed any of the others. Cluster stability was related to the homogeneity of the items included. Estimates of the proportion of nine-year-olds actually able to solve each item type (actually possessing the various skills) ranged from .41 to .89 (averages) or from .30 to .90 (estimates from individual analyses). These are shown in Tables 1 and 2.

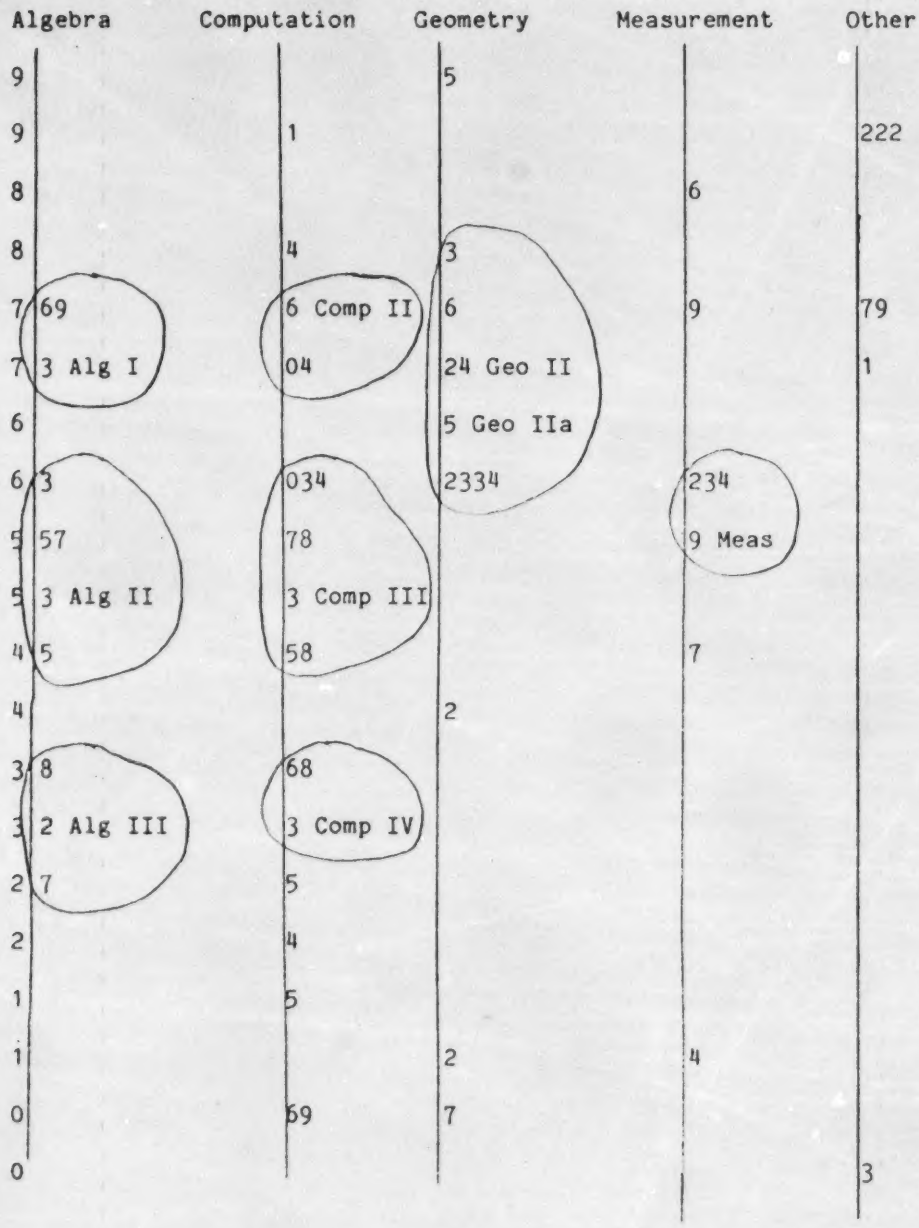
Table 2. Age 9: Summary Across Runs by Cluster

	Estimates of Proportion Able to Solve					Observed Item Difficulties		
Cluster	N	Mean	S.D.	Min	Max	1	2	3
Add	8	.89	.004	.89	.90	.863	.880	.914
Sub	5	.76	.007	.75	.77	.685	.721	.844
Place	5	.79	.009	.78	.80	.791	.796	.859
Count	5	.83	.021	.81	.86	.802	.920	.925
Units	6	.77	.058	.69	.84	.788	.808	.846
Geo	6	.69	.127	.50	.84	.536	.568	.771
SMD2	5	.42	.093	.30	.54	.187	.278	.340
Ruler	6	.53	.084	.41	.66	.367	.619	.659
Alg	6	.41	.070	.30	.50	.234	.240	.324

Results and Interpretation: Age 13

Stem-and-leafs of crude item difficulties for Age 13, Booklet 6 are displayed in Figure 3. As for Age 9, homogeneous sets are circled and labeled to show the clusters drawn from each.

Figure 3. Age 13 Booklet 6 Stem-and-Leafs of Crude Item Difficulties



In general, the Age 13 clusters are less homogeneous than Age 9 clusters. For this reason, less descriptive labels are used. In the following descriptions, the three exercise parts are described in the order of their appearance in Age 13 Booklet 6. These descriptions can therefore be aligned with the statistics tabled later in this section. Comp II included a free-response item asking what fraction of some marbles is blue (answer 5/7), a multiple-choice item requesting the denominator of 3/5, and another free response item asking for the quotient in 9.6 divided by 3. The Comp III and Comp IV clusters are the most homogeneous of any at Age 13. Comp III includes three free-response items requesting the square roots of 9, 49, and 25. Comp IV exercises were multiple-choice, requesting decimals equal to 1/4, 3/8, and 5/6. The "repetend" notation (e.g., $.\overline{77}$ for 7/9) appeared in the distractors for the second and third of these exercises, and in the correct response for the third. Alg I included two number line exercises, one free-response (Mark an X where 1.5 should be) and one multiple-choice (What number is at point A), and a multiple-choice item asking which number sentence could be used to solve a simple addition with one omitted addend. The Alg II cluster included two free-response items requiring solutions to simple equations and a multiple-choice "translation" item asking which equation in X and Y expresses the idea that when two numbers are added the order can be changed. Alg III included a multiple-choice exercise requiring the inference that $(-3768/n) = 314$ implies n is negative, a free-response item requiring the inference $f(n) = n + 5$ implies $f(3) = 8$, and a graphing exercise, put an X at the point (3,-2). The Geo II cluster consisted of three multiple-choice exercises requesting the number of corners, faces, and edges a cube has, given illustrations of each of these terms. In Geo IIa, the first exercise was a composite of four exercise parts, pre-scored. Exercise parts 32A, B, C, and D stated that two triangles were CONGRUENT, and then presented four

statements of attributes of congruent triangles as true-false questions (equal sides, equal angles, equal areas, and superposability). These were used to derive a single dichotomous variable, correct if all four statements were answered "true", else incorrect. This was to provide a single, more reliable measure of knowledge of the concept "congruent triangles". The remaining exercises in Geo IIa were multiple-choice, requiring the examinee to select line drawings representing a cylinder and a sphere. All three concern knowledge of geometric terms. The last cluster, Meas, includes multiple-choice questions on the number of quarts in a gallon, ounces in a pound, and feet in a yard. (Exercise numbers assigned to each of these items in Booklet 6 are shown in Table 3, discussed below.)

The 21* analyses run on these clusters are diagrammed in Figure 4. Interpretation of this figure is the same as for Figure 2, described in the last section. Four of the Algebra and Computation clusters at Age 13 proved to measure a single skill. As shown by analyses 6, 14, and 16, the Comp II, Comp IV, Alg II and Alg III clusters collapsed, leaving only six distinct skills. In spite of the greater heterogeneity of items within clusters at Age 13, there was only one analysis in which items within clusters were found to require distinct skills. This was analysis 1, involving the Alg I and Comp II clusters. In the final model for the six items in these two clusters, roughly 6 percent could solve only two of the Comp II exercises and one of the Alg I exercises, 5 percent could solve only the remaining two Alg I exercises, and 73 percent could solve all six exercises, as shown in Figure 4. This breakdown of the clusters in analysis 1 as well as the collapsing of Comp II, Comp IV, Algebra II, and Algebra III indicate that at Age 13 the Algebra vs. Computation distinction could not be sustained empirically.

*Age 13 analyses are numbered 1-7 and 9-22.

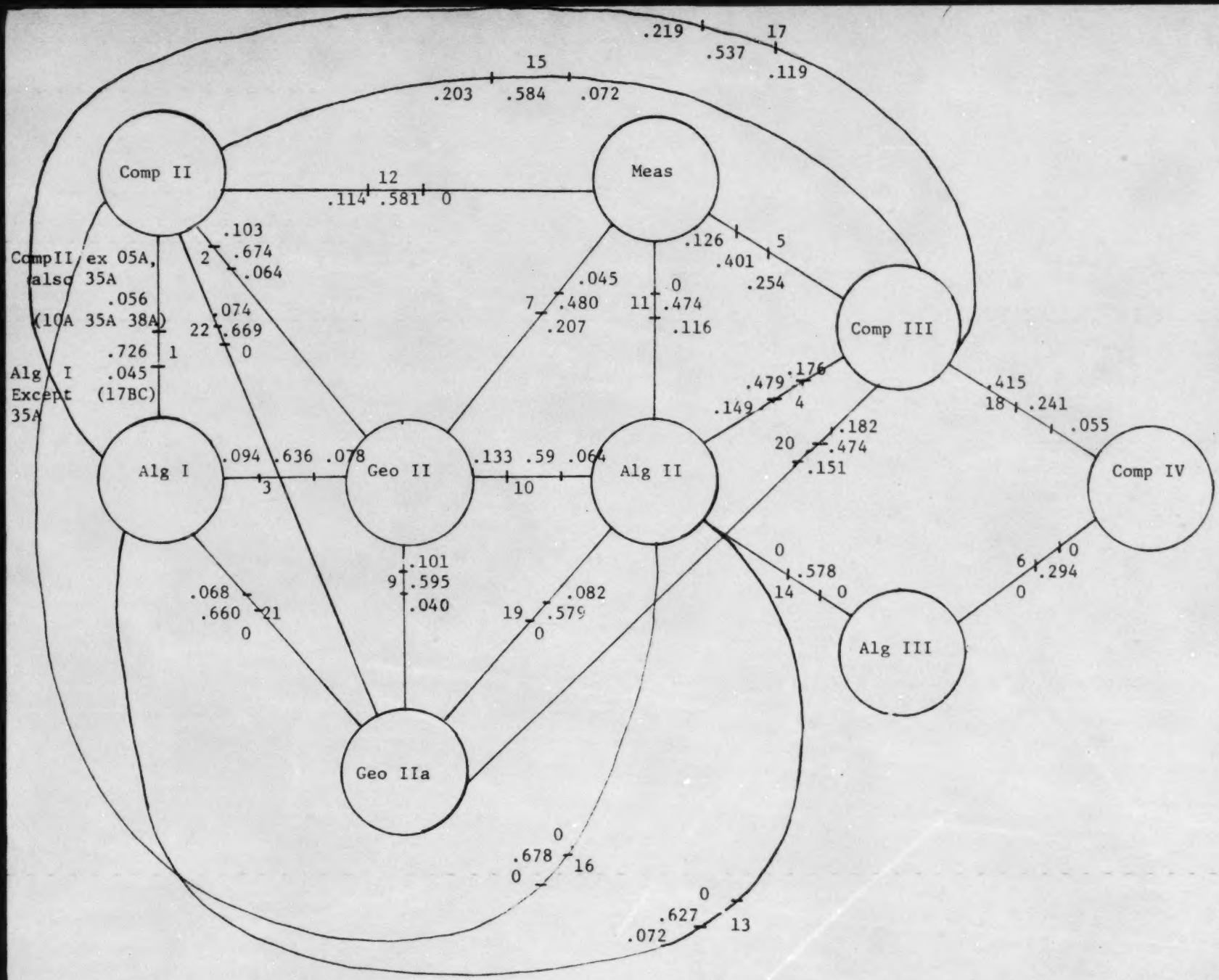


Figure 4. Age 13 Analyses and Proportions in Each Skill Pattern

Hierarchical relationships among Age 13 skills were as follows. The skills required for the Geo II, Geo IIa, Comp III, and Meas clusters could all be acquired independently of one another. Geo IIa was prerequisite to Alg I, and Geo IIa, Alg I, and Meas were all prerequisite to the Comp II-Comp IV-Alg II-Alg III skill. That is, no one could solve Comp II etc., items who could not also solve Geo IIa, Alg I, and Meas items, and no one could solve Alg I items who could not also solve items in Geo IIa. The only exception to these hierarchical patterns was in analysis 1, which indicated that 5 percent of the examinees could solve two of the three Alg I exercises but not two of the three Comp II exercises. As was noted for Age 9, hierarchical dependencies may arise from at least two sources: the logical dependency of more advanced content upon prerequisite or component skills, and the conventional structure of the curriculum, in which some skills are almost universally introduced earlier than others. The Alg I exercises involving number lines and a simple number sentence are probably subordinate to Alg II, Alg III, etc. for both of these reasons. The prior status of Meas (common unit conversions) and Geo IIa (definitions of "congruent," "sphere" and "cylinder") appears more related to curricular structure than to content.

Tabulations of item parameter estimates and estimates of proportions able to solve each item type appear in Table 3. It is formatted exactly like Table 1, in the last section. The leftmost columns give the cluster and item identifications (from Age 13 booklet 6), and p-values. To the right of each cluster appear three-column summaries from all the analyses in which that cluster appeared. These summaries include the analysis number and the name of the other cluster involved (first column, line 1), the chi square (first column, line 2) and the estimated proportion able to solve the cluster (first column, line 3), as well as estimated true positive and false positive rates for each item in the cluster (columns 2 and 3). The

Table 3. Age 13 Estimates of Misclassification Probabilities by Item

Cluster	Item	P	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP	Set/ Opp/P	TP	FP
Comp II	05A	.715	01:Alg I	.86	.34	02:Geo II	.86	.21	12:Meas	.87	.35	15:Comp III	.82	.32	16:Alg II	.88	.37	22:Geo IIa	.85	.31			
	10A	.769	59.38	.86	.46	56.23	.85	.49	62.77	.87	.53	73.10	.85	.46	67.40	.88	.53	65.80	.86	.49			
	38A	.751	.76* (.73,.78,.78)	.89	.25	.78	.87	.34	.70	.91	.40	.78	.90	.21	.68	.91	.41	.74	.89	.35			
Comp III	31A	.649	04:Alg II	.96	.05	05:Meas	.96	.05	15:Comp II	.96	.05	17:Alg I	.96	.05	18:Comp IV	.96	.05	20:Geo IIa	.96	.05			
	31B	.601	78.73	.91	.01	59.70	.91	.01	73.10	.91	.01	68.62	.91	.01	87.91	.91	.01	68.03	.91	.01			
	31C	.654	.66	.98	.04	.66	.98	.04	.65	.98	.04	.66	.98	.03	.65	.98	.03	.65	.98	.03			
Comp IV	36A	.378	06:Alg III	.90	.16	18:Comp III	.90	.16															
	36B	.259	55.86	.79	.04	87.91	.78	.04															
	36C	.256	.29	.83	.02	.30	.82	.02															
Alg I	17B	.779	01:Comp II	.92	.29	03:Geo II	.93	.37	13:Alg II	.93	.42	17:Comp III	.92	.35	21:Geo IIa	.94	.36						
	17C	.801	59.38	.91	.42	46.92	.92	.47	76.13	.92	.52	68.62	.93	.40	66.55	.92	.49						
	35A	.733	.77* (.77,.77,.78)	.83	.37	.73	.83	.46	.70	.85	.46	.76	.82	.47	.73	.83	.48						
Alg II	03A	.556	04:Comp III	.78	.18	10:Geo II	.76	.18	11:Meas	.80	.21	13:Alg I	.77	.19	14:Alg III	.79	.24	16:Comp II	.76	.13	19:Geo IIa	.76	.17
	08A	.532	78.73	.72	.22	32.92	.70	.21	57.45	.73	.24	76.13	.70	.25	58.83	.73	.27	67.40	.68	.22	52.66	.69	.22
	09A	.639	.63	.84	.30	.66	.83	.27	.59	.84	.35	.63	.85	.29	.58	.88	.31	.68	.81	.27	.66	.83	.27
Alg III	15A	.323	06:Comp IV	.42	.28	14:Alg II	.35	.28															
	23A	.282	55.86	.39	.24	58.83	.36	.18															
	30B	.391	.29	.54	.33	.58	.55	.17															
Geo II	37A	.843	02:Comp II	.95	.55	03:Alg I	.95	.56	07:Meas	.96	.59	09:Geo IIa	.96	.58	10:Alg II	.95	.56						
	37B	.750	56.23	.93	.24	46.92	.94	.28	52.84	.95	.32	61.00	.95	.29	32.92	.94	.25						
	37C	.727	.74	.91	.20	.71	.92	.24	.69	.94	.26	.70	.93	.27	.72	.91	.24						
Geo IIa	32ABCD	.345	09:Geo II	.48	.11	19:Alg II	.50	.13	20:Comp III	.49	.10	21:Alg I	.47	.10	22:Comp II	.47	.09						
	33A	.782	61.00	.96	.48	52.66	.94	.56	68.03	.95	.51	66.55	.93	.50	65.80	.93	.49						
	33C	.639	.64	.82	.32	.58	.87	.32	.62	.84	.30	.66	.83	.27	.67	.83	.26						
Meas	16B	.633	05:Comp III	.84	.40	07:Geo II	.84	.40	11:Alg II	.85	.43	12:Comp II	.82	.38									
	16C	.622	59.70	.91	.30	52.84	.86	.35	57.45	.93	.35	62.77	.88	.27									
	16D	.652	.53	.87	.40	.53	.91	.36	.47	.90	.43	.58	.85	.38									

*Average proportion for 3 items. In this run, the Alg I and Comp II items showed distinct skill profiles within sets.

stability of the true positive and false positive estimates is closely related to the stability of the estimated proportion able to solve the cluster. These latter estimates for each cluster are summarized in Table 4, which also repeats the observed p-values for each item. For all the Age 13 clusters except Alg I, estimated proportions able to solve were as high or higher than average item difficulties, indicating that random guessing is relatively rare among 13-year-olds. (Guessing on multiple-choice items increases p-values, so that average item difficulties exceed the proportion possessing the skill.) Lower true positive rates for exercises at this age level indicate that more of the individual exercises examined present unique difficulties or require specific information not shared with other items in the cluster. Thus, examinees possessing the skills these items require in common may nonetheless err on one or another of the separate exercises.

Table 4 shows that estimated proportions able to solve are not in close agreement for exercises in the four clusters that collapsed to a single skill. They range from .30 (Comp IV) through .44 (Alg III) and .63 (Alg II) to .74 (Comp II). The structure of the set of Age 13 analyses implies that for at least one of these clusters, estimates from different analyses must have been quite variable, and, in fact, the problem appears to be with the Alg III cluster alone (see standard deviations in Table 4). It can be seen in Figure 4 that Alg III was used in only two analyses, 14 (with Alg II) and 6 (with Comp IV). In each of these analyses, a two-class model gave an acceptable fit, indicating a single common skill for all three clusters. However, the proportion able to solve Alg II and Alg III was estimated as .578, while the proportion able to solve Comp IV and Alg III was estimated as only .294. Additional analyses would be required to probe the reasons for this anomoly. In particular, it would be useful to analyze Alg II and Comp IV together. Runs for analysis sets 6 and 14 in which the proportion possessing the common skill was constrained to equal a constant would also be informative. Work with other item sets

Table 4

Age 13: Summary Across Runs by Cluster

	Estimates of Proportion Able to Solve				Observed Item Difficulties		
Cluster	N	Mean	S.D.	Min	Max	1	2	3
Comp II	6	.74	.042	.68	.78	.715	.751	.769
Comp III	6	.66	.005	.65	.66	.601	.649	.654
Comp IV	2	.30	.007	.29	.30	.256	.259	.378
Alg I	5	.74	.028	.70	.77	.733	.779	.801
Alg II	7	.63	.037	.58	.68	.532	.556	.639
Alg III	2	.44	.205	.29	.58	.282	.323	.391
Geo II	5	.92	.013	.91	.94	.727	.750	.843
Geo IIa	5	.84	.019	.82	.87	.345	.639	.782
Meas	4	.88	.028	.85	.91	.622	.633	.652

indicates that the two-class model may be relatively insensitive to such constraints, i.e., that the chi square may increase only slightly if different latent class proportions are fixed in advance. (This has not been found for more complex models.) Another possible explanation is found in the decision rules for the series of runs for each analysis. In analyses 6 and 14, two-class chi squares of 56 and 59, respectively, were obtained (see Table 3). Thus, no additional runs were performed on these item sets. The introduction of a third latent class might have resulted in substantial reduction of these (already nonsignificant) values leading to new "final" models in which Alg II, Alg IV, and Comp IV were hierarchically related. This would resolve the discrepancies among Alg II, Alg III, Comp II, and Comp IV. Further work on the decision rules is called for. Hierarchical organization among Alg II, Alg III, and Comp IV is plausible, and would extend the hierarchical relationship observed between Alg I and Alg II. This interpretation, that the marginally satisfactory 2-class fits of analyses 6 and 14 should not have been accepted, is also supported by the nearly hierarchical organization of the Comp II, Comp III and Comp IV clusters shown in analyses 15 and 18. The fact that only 6 percent of the examinees could solve Comp IV and not Comp III and only 7 percent could solve Comp III and not Comp II but percentages in the other direction are 41 percent and 20 percent makes it seem unlikely that Comp II and Comp IV require the same skills.

In summary, the Age 13 analyses showed that four of the nine clusters, Comp II, Comp IV, Alg II, and Alg III required a single skill. This finding is suspect, however, on two grounds. First, a nearly hierarchical relationship was found among Comp II, Comp III, and Comp IV. Second, the Alg III cluster was included in only two analyses, which yielded quite different estimates of the proportion able to solve Alg III items. These were two of the three runs indicating that Alg II, Comp II, Alg III, and

Comp IV required a common skill. Reconsideration of analyses 6 and 14 would require changing the a priori decision rules which governed all the Age 9 and Age 13 analyses reported. However, further investigation of these findings is called for. Assuming that further runs would indicate a hierarchical relationship among Alg II, Alg III, and Comp IV, the equivalence established between Alg II and Comp II would still stand. Thus, no more than 8 skills would be required to explain item response patterns in these data, with the great majority of 13-year-olds mastering first Meas and Geo IIa, then Alg I, and then progressing to Alg II-Comp II, then on to Alg III and Comp III, followed by Comp IV. Geo II does not appear in the hierarchy but is roughly parallel to Geo IIa in difficulty and time of acquisition. Estimates of the proportion of examinees possessing these skills ranged from .30 (Comp IV) to .92 (Geo). These are shown in Table 4.

Results and Interpretation: Age 17

At Age 17, a different analytic strategy was followed. Rather than mapping relationships among homogeneous item triples in a single booklet, sets of six exercises were drawn from each of six separate booklets. All these 36 items focussed on a common, broadly conceived skill, and no two were parts of the same exercise. Rigid rules were not followed in carrying out these analyses. Rather, each series of runs was pursued until all additional latent classes suggested by examination of residuals had been tried. Of course, no latent classes were retained unless their introduction resulted in a statistically significant improvement in the fit of the model.

There were several reasons for adapting a different strategy at Age 17. First, patterns of findings at Age 13 suggested that it might be premature to stop with the first run yielding a nonsignificant chi square.

Second, the large number of distinct skills found at both Age 9 and Age 13 suggested that the skills obtained might be of limited generalizability. It had been anticipated that fewer, broader skills would be identified. In order to identify broader skills, the six-item sets at Age 17 were each formed from six exercise parts that appeared to require the same common concept, that letters can replace numbers in statements of equations and inequalities. (At Ages 9 and 13, three items for each of two concepts had been used in each set.) To minimize common method variation that might give rise to spurious skill distinctions, only the "A" parts of multi-part exercises were included. Replication across six separate examinee samples (assured by sampling of exercises from six different booklets) also promised to increase generalizability, and permitted significance testing to compare estimates of latent class proportions from different exercise sets, since these estimates could be assumed statistically independent.

To identify the exercise sets examined at Age 17, the Age 17 appendices were examined and all exercises were noted that appeared to require an understanding that in that exercise, letters took the place of numbers. The 88 exercise parts identified (from 75 distinct exercises) included some classified as algebraic manipulation (solving equations, simplifying and factoring, plotting, graphs), mathematical skills and computation, numbers and numeration, understanding, translation, and other topics. Actual exercise parts were then examined on microfiche, and roughly 20 were rejected as not relying on the common skill. This left just six of the twelve Age 17 booklets with at least six acceptable exercises. Six exercises were drawn from each of these booklets, as shown in Table 5. The smaller number of six-item sets analyzed at Age 17 does not imply a lower commitment of either time or computer resources. More distinct exercise parts are involved at Age 17 than at either Age 9 or Age 13, it was necessary to access more tape files because multiple booklets were involved,

Table 5

Exercises Examined in Age 17 Analyses, p-values, and Latent Classes Required

				Latent Classes*			
Set	Booklet	Exercise	P	NULL	REAS	INT	ALL
1	01	05A	.233	0			1
		09A	.568	0			1
		22A	.041	0			1
		24A	.706	0			1
		33A	.662	0			1
		39A	.796	0			1
	x^2_{50} = 59.29.	Proportions =		.574			.426
2	02	02A	.725	0	0	1	1
		04A	.048	0	0	0	1
		09A	.449	0	0	0	1
		18A	.702	0	1	1	1
		27A	.883	0	1	1	1
		30A	.473	0	1	0	1
	x^2_{48} = 64.33.	Proportions =		.394	.066	.083	.457
3	03	03A	.521	0	1	1	1
		14A	.174	0	0	0	1
		18A	.788	0	1	1	1
		23A	.581	0	1	1	1
		35A	.187	0	0	0	1
		38A	.262	0	0	1	1
	x^2_{48} = 52.37.	Proportions =		.582	.104	.057	.256
4	04	05A	.124	0	0	0	1
		13A	.580	0	1	1	1
		20A	.490	0	1	0	1
		26A	.686	0	1	1	1
		29A	.075	0	0	0	1
		35A	.578	0	1	1	1
	x^2_{48} = 63.08.	Proportions =		.553	.112	.152	.184
5	07	07A	.343	0	0		1
		17A	.370	0	0		1
		24A	.531	0	0		1
		28A	.526	0	1		1
		33A	.470	0	1		1
		38A	.193	0	0		1
	x^2_{49} = 62.08.	Proportions =		.565	.077		.358

* A "1" indicates that examinees in that latent class could solve the corresponding item; "0" indicates that they could not.

Table 5 (Cont.)

Exercises Examined in Age 17 Analyses, p-values, and Latent Classes Required

				Latent Classes*			
Set	Booklet	Exercise	P	NULL	REAS	INT	ALL
6	09	05A	.178	0	0		1
		06A	.382	0	0		1
		19A	.368	0	1		1
		20A	.392	0	0		1
		21A	.561	0	1		1
		33A	.629	0	1		1
	x^2_{49} = 93.23.		Proportions =	.520	.061		.419

and most important, the absence of distinct, homogeneous clusters made actual fitting more difficult.

Results of the six analyses will be discussed in order of increasing complexity. Set 1 yielded a satisfactory fit to the simplest, two-class model. The chi square and estimated latent class proportions for this and other analyses are displayed in Table 5. It should be noted that exercise p-values for set 1 ranged from .04 to .80, which makes it appear unlikely that a single skill distinction could explain examinee performance across the entire set. In fact, the true positive rate for exercise 22A (with a p-value of .04) was estimated as only .10. It may be more reasonable to think of this item as requiring some additional skill not required by the remaining five items in the set, rather than supposing that it requires the same skill but that those possessing the requisite skill had only one chance in ten of answering it correctly. These two interpretations are, in fact, interchangeable. They are statistically equivalent, yielding identical predicted response pattern frequencies, residuals, and chi squares. The more reasonable model, in which item 22A alone requires some additional skill, would include three latent classes, for examinees who could solve none of the exercises, all but 22A, and all of the exercises. This model, however, is said to be not identified. Its parameters cannot be estimated because they cannot be uniquely determined. In particular, there is no way, with this model, to distinguish a student in the second class who gives a false positive response to item 22A from a student in the third class who gives a true positive. By making compensating adjustments in the latent class proportions and item 22A's misclassification probabilities, identical predicted response pattern proportions can be obtained. The two-class model actually fitted to set 1 represents one possible set of parameter estimates for the nonidentified three-class model. Another set could be obtained by fixing the true positive rate for item 22A at 1.00, in which case latent

class proportions would be .574 (null), .385 (all but 22A), and .041 (all items). Fixing the second latent class proportion at any value between 0 and .385 or fixing the true positive rate for 22A at any value between .096 and 1 would yield an equivalent model. In conclusion, the analysis for set 1 might be interpreted as showing a single skill distinction with an extremely low true positive rate for one item, or as showing that all items except 22A require the same skill and the latter requires a unique additional skill.

The next two sets, in order of complexity, were sets 5 and 6, for which the final models included three latent classes. For set 6, the final model fit relatively poorly, as indicated by a chi square of 93.23 on 50 degrees of freedom. No additional classes could be found which reduced this value. As shown in Table 5, the skill distinctions found in sets 5 and 6 would not be predicted on the basis of item p-values. While any interpretation of these skill distinctions would require cross-validation, it appears in each case that examinees in the intermediate state can solve just those items that could be answered correctly by reasoning, without formal training in algebra. These are non-routine problems, multi-step word problems, and items testing number and numeration concepts in a form different than that typically encountered in instruction, and may be more related to general intelligence and less to schooling than the other items in sets 5 and 6. No explanation is offered for the relatively poor fit of set 6.

The remaining sets all required four latent classes for an adequate fit. Sets 3 and 4 are treated first, because the latent classes in these runs conformed to a hierarchical skill pattern. As shown in Table 5, for set 3 examinees could solve all of the items, all but 14A and 35A, all but 14A, 35A and 38A, or none of the items. A similar pattern emerged for set 4. As for sets 5 and 6, skill distinctions appear to reflect the solubility of items by reasoning as opposed to formal training. Examinees

in intermediate states can solve items concerning abstract properties of the number system, especially multiple-choice exercises, but fail to solve routine computational exercises that appear to depend upon specific, formal training in algebra. Unlike sets 5 and 6, the patterns of intermediate states in these two sets correspond precisely to the patterns indicated by item difficulties. In set 3, for example, 10.4% can solve only items with p-values above .5, an additional 5.7% can solve all items with p-values above .2, and 25.6% can solve all items.

Set 2 yielded the most complex pattern of latent classes. Since the intermediate states are not nested hierarchically, data from set 2 are not compatible with the assumption of an ordered, unidimensional continuum of content knowledge or content acquisition. It may be no accident that set 2 is also the easiest set of items. In other work with reading comprehension item response data, the author has found similar anomolies with very easy items. It is tempting to assume that these items are amenable to solution by several strategies. Some students have a strategy that works for items 02A, 18A, and 27A, while others have a strategy for 18A, 27A, and 30A. Unfortunately, it is not clear what these strategies would be, and it appears unlikely that many researchers would predict in advance that these two intermediate classes would be the ones to emerge. The pattern of latent classes here is not consistent with patterns of item difficulty.

Despite the variety of models fitted to these six item sets, some important consistencies do emerge. Note first the estimated proportions in the NULL classes for the six runs, those unable to solve any of the items. With the exception of the anomolous value for the set 2 (the last set just discussed), these estimates fall in the narrow range from .520 to .582. Even the conservative procedure of testing the difference between the smallest of these and the largest yields a non-significant t of 1.86.

This indicates that the NULL proportion was consistent across sets 1 and 3-6, within sampling error. As discussed above, for the sets in which intermediate classes emerged, one category typically appeared to be of students who could solve non-routine or multiple-choice exercises but not conventional algebra exercises. For each of sets 2-6, the intermediate class that appeared to best represent this pattern was identified. In Table 5, patterns for this class appear in the column headed "Reas". Estimated proportions for this class ranged from .061 to .112. As for the NULL proportion estimates, the conservative t-test of the smallest against the largest of these estimates was calculated. A non-significant value of 1.27 was obtained, indicating that the assumption of a common skill pattern, detected across analyses, could not be rejected. Roughly 8 percent of examinees could correctly answer items soluble by reasoning, but not those requiring formal training in algebra. Further investigation of the course-taking patterns of these students is clearly warranted.

The final commonality observed across these analyses concerns the unidimensionality of the skill continuum. In all but set 2, latent classes showed a Guttman scale pattern. As noted earlier, the anomoly in set 2 was for items markedly easier than those in the other five sets. Thus, it may be concluded that a single continuum of content acquisition underlies all moderately difficult to difficult exercises involving the use of letters to represent real numbers. As with the hierarchical relationships detected at ages 9 and 13, this continuum probably results from both the logical structure of the exercise content and the conventional structure of the mathematics curriculum in American schools.

Conclusions

Developing appropriate, sensitive methods to analyze the National Assessment data is a difficult challenge. If analysis and reporting are to go beyond the level of individual items, new methods must be developed or existing methods adapted to a matrix-sampled data base in which exercises are conceived not as measures of a few common, underlying traits, but as representing many distinct classes of specific performances, each of interest in its own right. In other words, individual NAEP exercises represent more or less distinct content domains, each of which was specified because it was of some intrinsic interest. Classical test theory, with its focus on the detection of stable individual-difference variation against a background of measurement error, is ill-suited to the analysis of content- or domain-referenced instruments. Newer item response theory (IRT, or latent trait) models are as yet only practical when applied to large numbers of items that may be assumed unidimensional. The large number of distinct, independent skills detected in this investigation strongly suggest that IRT models will be of limited value.

Both classical and IRT models begin with the simplifying assumption that items measure a common underlying skill - an assumption of unidimensionality. In this study, latent class models rather than latent trait models were used, and a different simplifying assumption was involved -- that examinees may be classified as to whether they can or cannot solve each individual item, and that these two possible classifications correspond to just two distinct probabilities of a correct response to the item. Other assumptions, especially conditional independence, are the same for latent class and IRT models. The latent class models give considerable flexibility in accounting for the specific characteristics of individual items, and

also permit the representation of multidimensional structures of skill possession, or non-linear patterns of skill acquisition.

This study has demonstrated the utility of latent class models in describing NAEP data, but has also highlighted technical problems in need of further investigation. Significance testing when data are from a stratified cluster sample is an unsolved problem. The design effect adjustment used in this study is clearly not the best possible. The only estimates presently available are maximum likelihood, and while some large-sample properties of these estimates are known, further work on the problem of estimation is called for. Even more important, work is needed on the problem of model selection. The methods developed in this study for the analysis of residuals are of considerable help in finding the latent classes required to represent a given set of response patterns. However, these methods cannot as yet be automated, and are far from infallible. Beyond the problem of modeling individual exercise sets is the broader problem of selecting exercise sets for analysis. The strategy followed at ages 9 and 13 probably yielded clusters that were too tight, sharing more common method variance than desirable. The strategy used at age 17, on the other hand, yielded broader exercise clusters and facilitated replication across booklets, but did not yield as comprehensive a map of exercise skill requirements. Extension of computational methods to accommodate larger numbers of items would help to solve this problem.

Despite present limitations of the methodology, some important generalizations do emerge from the analyses reported. At age 9, roughly 89 percent of the children could do simple addition problems correctly. Subtraction, counting, place value problems, unit conversions, and simple geometric terms and concepts were all relatively independent of one another, i.e., required, for the most part, distinct skills. These all tended to be acquired after the addition skill. Roughly 77 percent of nine-year-olds

could solve problems of these kinds. More difficult were two-digit subtraction, multiplication, and divison problems, number line and number sentence problems, all available to roughly 41 percent of the population, and hierarchically dependent upon the easier skills. The skills required to use a ruler were only slightly easier (53 percent) and also tended to be acquired after the less difficult skills.

At Age 13, several levels of Algebra/Computation skills were detected, but were not clearly distinguished. The skills required to work with number lines and number sentences, to express part of a small set as a common fraction, and to do a simple long division problem were available 74 percent of the sample. Sixty-six percent could interpret the radical sign, but only 30 percent could give the decimal equivalents of common fractions. These skills were not strictly hierarchical, but only a few percent of the examinees possessed the easier skills who lacked the more difficult ones. The skills required for a variety of algebra exercises (including topics in numbers and numeration) were available to 40 to 60 percent of the examinees. As with computation, these skills were nearly hierarchical. Relatively independent skills were required for simple unit conversions (possessed by 88 percent of the examinees) and geometry facts and concepts (84 and 92 percent, respectively).

The Age 17 analyses were structured differently, and did not yield a comprehensive map of skills possession. Only algebra exercises were examined, but these were drawn from six different booklets. It was found that 52 percent of 17-year-olds were unable to solve any items in which letters represented variable numerical quantities. Another 8 percent could solve problems of this kind that did not appear to depend heavily upon formal training in algebra, but were unable to solve more routine algebra problems. For all but the easiest exercises, data were compatible with the assumption of a linear skill hierarchy.

The Age 17 strategy for planning exercise sets could be fruitfully

applied to other broad topics. Further work with these models and methods should increase our understanding of the academic capabilities of American youth.

References

Bock, R.D. & Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Dayton, C.M. & Macready, G.B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Goodman, L.A. A new model for scaling response patterns: An application of the quasi-independence concept. Journal of American Statistical Association, 1975, 70(352), 755-768.

Haertel, E.H. Determining what is measured by multiple choice tests of reading comprehension. Unpublished doctoral dissertation, University of Chicago, 1980.

Lazarsfeld, P.F. & Henry, N.W. Latent structure analysis. New York: Houghton Mifflin, 1968.

Proctor, C.H. A probabilistic formulation and statistical analysis for Guttman scaling. Psychometrika, 1970, 35, 73-78.